DOCUMENT RESUME

ED 189 163 TM 800 334

AUTHOR Wainer, Howard: Wright, Benjamin D.
TITLE Robust Estimation of Ability in the Rasch Model.
INSTITUTION Bureau of Social Science Research, Inc., Washington, D.C.
SPONS AGENCY Law Enforcement Assistance Administration (Dept. of Justice), Washington, D.C.
PUB DATE Apr 80
GRANT 78-NI-AX-0047
NOTE 55p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Error of Measurement: Guessing (Tests): Item Analysis: *Latent Trait Theory: *Scoring: Simulation: *Statistical Bias: Testing Problems
IDENTIFIERS *Jackknife (Statistics): Monte Carlo Studies: *Rasch Model

ABSTRACT

The pure Rasch model was compared with four modifications of the model in a number of different simulations in order to ascertain the comparative efficiencies of the parameter estimations of these modifications. Because there is always noise in test score data, some individuals may have response patterns that do not fit the model and their abilities cannot be estimated accurately. The modifications used to overcome this difficulty include: the traditional correction for guessing, jackknifing, a robustified jackknife, and a method (WIM) based on an analysis of residuals. The simulation data included a guessing model with four characteristics: the need to guess, the invitation (a function of the item), the inclination (a function of the person), and glitch. The simulations were also varied by including: 3 difficulty levels, 3 test length, 2 levels of width, either 2 or 5 choice items, and 4 candidate ability levels. Both the accuracy of parameter estimation and the sampling error were combined to obtain indices of the relative efficiency of the five models. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189163

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ROBUST ESTIMATION OF ABILITY IN THE RASCH MODEL[1]

Howard Wainer
Bureau of Social Science Research
1990 M Street, N.W.
Washington, D.C. 20036

and

Benjamin D. Wright
The University of Chicago
Chicago, Illinois 60637

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. WAINER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ABSTRACT

Estimating ability parameters in latent trait models in general, and in the Rasch Model in particular is almost always hampered by noise in the data. This noise can be caused by guessing, inattention to easy questions, and other factors which are unrelated to ability. In this study several alternative formulations which attempt to deal with these problems without a reparameterization are tested through a Monte Carlo simulation. It was found that although no one of the tested schemes is uniformly superior to all others, a robustified Jackknife stood out as the best one in general, it was also super efficient for tests with forty or fewer items. It is proposed that this sort of Jackknifing scheme for estimating ability be implemented for practical work.

---

[1] This research was funded through a grant from the Law Enforcement Assistance Administration (78-NI-AX-0047) to the Bureau of Social Science Research, Howard Wainer, Principal

TM 800334

Investigator. We would like to thank Ronald Mead, Anne Morgan and James Ramsay for kind, generous, and invaluable help at various stages of the project.

1. Introduction and Background

Latent trait models as a class, and the Rasch model in particular, have begun to have substantial impact on the construction and scoring of mental tests. Through the use of latent trait models, measures of individual ability as well as item difficulty can be obtained that have important practical and statistical properties. For example, if the Rasch model fits, the measures of ability and difficulty obtained are interval scaled thus making the quantitative study of change possible. The Rasch model characterization of a person's performance on an item as a function of the difference between that person's ability and the difficulty of the item yields the useful result that one can obtain sample-free item calibration as well as test-free person measurement. There are many more reasons why a latent trait formulation is an important one (see for example, Rasch, 1960; Wright, 1968; Lord and Novick, 1968; Wright and Panchapakesan, 1969; Bock and Wood, 1971; Wright, 1977; Hambleton et al, 1978; Wainer, Morgan and Gustafsson, 1979).

The problem in harvesting the benefits of latent trait models is the problem of fit. These benefits only follow when the model fits. Studies of robustness (Lord and Novick, 1968, p 492) indicate that certain parameters are robust with respect to modest deviations from the underlying assumptions; in particular it seems that the Rasch model yields rather good estimates of ability and difficulty even when its assumption of equal slopes is only roughly

approximated. The models which parameterize differential slopes have difficulty recovering the slope parameters even when the data do fit their model. This is not a topic of this paper. We merely want to indicate that attempts to expand the one parameter model to encompass additional possible characteristics of the data through an increase in the number of item parameters does not appear to be completely successful yet. Slope parameters are not well estimated in testing situations with only a few hundred individuals (Lord, Reference Note 1), and lower asymptotes, introduced to deal with guessing, cannot be consistently estimated at all (Ree and Jensen, Reference Note 3).

II. The Problem

The Rasch model has many practical benefits if it fits. It can never fit exactly however, because there are always disturbances. These disturbances often take the form of; guessing (a person of low ability gets a difficult item correct) and sleeping (a person of high ability gets an easy item wrong) (Wright and Mead, 1977). The model has a certain amount of robustness with respect to such aberrations, but they can make the estimation procedures both biassed and inefficient. The problem, then, is how to estimate the parameters of interest accurately and efficiently even when the data don't fit the model.

III. Some Choices

As a way to deal with this problem we shall consider five different estimation schemes. We shall compare these

alternatives over a variety of simulations. We shall assume that item difficulties are available and that all that is to be estimated are person abilities. This is a reasonable assumption because we can increase the calibration sample greatly, winnow from it individuals who have unusual patterns of response and so get a sub-set of individuals who are not 'noisey'. These individuals can then be used to get good estimates of item difficulty. However, the dual is not true, we cannot give a test of great length, and when we are reporting on persons we cannot eliminate individuals who do not behave exactly as the model dictates. We need to do our best to estimate abilities for everyone. Our task is to explore various estimation methodologies which assume the availability of item difficulties and try to estimate ability as accurately and efficiently as possible. It may be that some of the techniques we describe will be of some use in the estimation of item difficulties as well, but this is not our primary motivation.

The Rasch Model

$$P_{ij}=\exp(a_i-d_j)/[1+\exp(a_i-d_j]$$

Where P_{ij} is the probability of person i getting item j correct

and

a_i is the ability of person i (i=1,...,N), and

d_j is the difficulty of item j (j=1,...,L)

Scheme 1 - Pure RASCH

This is the standard maximum likelihood method for estimating Rasch abilities given a vector of item difficulties. It relies on the Rasch model property that raw score is a sufficient statistic for ability. Each raw score has a distinct ability associated with it. To find what it is we solve the equation shown in (1) for a_i; usually through Newton-Raphson.

$$r_i - SUM_j[p_{ij}] = 0$$

Or (1)

$$r_i - SUM_j[\exp(a_i-d_j)/(1 + \exp(a_i-d_j))]=0$$

where r_i is the raw score for person i.

Scheme 2 - Traditional Correction for Guessing

The traditional guessing correction is to assume that if a person does not know the answer to a question and guesses, than the probability of guessing correctly is 1/M, where M is the number of choices. Thus if we have an M choice test and an individual has C wrong we assume that he has an additional C/(M-1) correct that he guessed on. This is a crude attempt to put a lower asymptote on the item characteristic curve.

Scheme 3 - Jackknife

The Jackknife is an estimation scheme which was developed to reduce bias, and has been shown (Tukey, 1958) to be useful for hypothesis testing as well. The way that it works in our application is to construct a matrix of abilities A which has L-1 raw scores labelling the rows, and

L+1 columns. The first column, with elements A(r,1) are the abilities associated with raw score r, calculated through the method described in Scheme 1. The second column are the abilities based upon a test with the first item omitted. This test has only L-1 items. Each succeeding column represents abilities estimated through Scheme 1 but with that item omitted. Thus the kth column is a test of length L-1 containing all items except item k-1.

The Jackknifed pseudovalues of ability are:

$$a_j^* = LA(r,1) - (L-1)[x_j A(r-1,j+1) + (1-x_j)A(r,j+1)] \quad (2)$$

Where $x_j = \begin{cases} 0 \text{ if item } j \text{ is answered incorrectly} \\ 1 \text{ if item } j \text{ is answered correctly.} \end{cases}$

and the Jackknifed estimate of ability, a*, is just the mean of these a_j^*'s.

$$a^* = SUM_j[a_j^*/L] = LA(r,1) - [(L-1)/L]$$
$$SUM_j[x_j A(r-1,j+1) + (1-x_j)A(r,j+1)]$$

for J=1,L.

For reasons that will become clear when we discuss the results of the simulations, it is important that we notice that the Jackknifed ability estimates are easy to compute. For any test all one has to do is to compute the matrix A and then for each person to run across the matrix at that subject's raw score adding up the entries in that row for each item that is incorrect and jumping up one row for each item that is correct. Jumping occurs because when an item is correct the raw score for that person excluding that item is one less.

Next, there are two aspects of an estimator that concern

us. The first is that it reduces bias, i.e. the effects of odd response patterns. The Jackknife was developed as a method to reduce bias (Quenouille, 1956), so we have hopes that it will serve this purpose. Secondly, we would like an estimator that does not jump around too much with minor disturbances in the response vector. This quality has been termed 'resistance' (Tukey, 1977), and corresponds to an estimator having a sampling distribution with a small variance. The Jackknife is known to be modestly 'resistant' and so this quality is likely to be met in practice as well. Let us see how estimation with the Jackknife works.

Consider a test with ten items whose difficulties are uniformly distributed and span a range of four logits. These difficulties are shown below:

-2.00	-1.56	-1.11	-0.67	-0.22
0.22	0.67	1.11	1.56	2.00

This yields the raw score-to-ability transformation matrix A, shown in Table 1.

Insert Table 1 About Here

Consider how one would estimate the ability for a response vector of (1111110001). The raw score is 7 and so we sum the first six values associated with a raw score of 6 (since the first six items were correct). Next we add on the three values (associated with items 7, 8, and 9) associated with a raw score of 7 since these items were incorrect and so omitting them still yields a raw score of

7. Last we add on 0.68 the ability pseudovalue associated with a raw score of 6 for item 10 omitted. Summing these we obtain a total of 11.63. Next we multiply by 9/10 [(L-1)/L] and subtract from 11.50 [L x 1.15] yielding a Jackknifed estimate for this person's ability of 1.03. Referring back to Table 1 we see that a raw score of 6 yields an ability estimate of .56 which would have been the result if we treated this person's getting the last item correct as a wild guess and changed it to incorrect. On the other hand if we fully believed this response his raw score would have been 7 and his ability estimate 1.15. The Jackknife weighs these two extremes and places the estimate between them.

Next suppose that the response vector was (1111110010). Then we find that the pseudovalue of .68 associated with getting item 10 correct is replaced with .73 (for item 9) and 1.45 is replaced by 1.37. The net result of this changes the Jackknifed estimate from 1.03 to 1.06. This is just what a sensible person would do, since the second response pattern is more likely to have arisen through "proper" test taking, and so indicates a somewhat higher ability.

It appears that the Jackknife does what we want, although how well is yet to be determined. We do get the feeling however from this demonstration that the variance of the sampling distribution of the Jackknifed ability is apt to be small since wild disturbances in response pattern do not cause wild variations in the ability estimates. To see this

note that the ability estimate associated with the pattern (0111111001) is 1.09. We leave it to the reader to try other patterns to develop an intuition as to how this estimation scheme behaves. The Jackknife is not insensitive to response pattern (as Rasch estimates are) but does not jump around much. This will be demonstrated in the results section.

Scheme 4 - AMT-Robustified Jackknife

The pseudovalues obtained from Jackknifing suggest an additional estimation methodology. Consider the response pattern(1111110001) again. If we calculate the pseudovalue associated with each item using Equation 2 we obtain:

Item	Pseudovalue
1	1.42
2	1.51
3	1.69
4	2.05
5	2.41
6	2.95
7	-3.17
8	-2.36
9	-1.55
10	5.38

The mean of these pseudovalues yields the Jackknifed estimate of ability. Now consider these pseudovalues , and how they are combined in the Jackknife. There are two kinds of pseudovalues ; negative ones associated with incorrect responses, and positive ones associated with correct responses. The Jackknife could be understood as first averaging the negative ones, and so coming out with an average ability estimate based upon items missed, then averaging the positive ones for an ability estimate from the

items gotten right, and then combining those two averages, weighted by their sample sizes to yield the final Jackknifed estimate. We know that the mean can be a poor way to estimate location. In some situations (Andrews et al, 1972) it is the worst of all choices. Since we are concerned about unusual situations, perhaps the performance of the Jackknife can be improved through the choice of an estimator of location more robust than the mean.

Suppose we calculate the median of the positive pseudovalues. This is 2.05. The median of the negative pseudovalues is -2.36. Weighting these by seven and three respectively, summing and dividing by ten yields an estimated ability of .73. Whether or not this is better than the Jackknifed value of 1.03 is hard to say, but it is certainly in the ballpark.

Insert Figure 1 About Here

One of the winners of the Princeton Robustness Study (Andrews et al, 1972) was the Sine M-estimator (the AMT). This estimator has an influence function that is nearly like that of the mean for observations in close but goes to zero at the extremes. This implies that it will be efficient for nearly Gaussian distributions and robust against fat tails and outliers.

To understand how the AMT is calculated consider that in 'regular cases', likelihood estimation of the location and

scales parameters THETA, and SIGMA of a sample from a population with known shape leads to equations of the form:

$$\text{SUM}_j\ [-f'(z_j)/f(z_j)] = 0,$$

and

$$\text{SUM}_j\ [z_j f'(z_j)/f(z_j) - 1] = 0,$$

where f is the density function and $z_j = (x_j - \text{THETA})/\text{SIGMA}$.

M-estimates of location are solutions, T, of an equation of the form

$$\text{SUM}_j\ \text{PSI}[(x_j-T)/s] = 0$$

where PSI is an odd function and s is estimated either independently or simultaneously.

The Sine M- Estimate (AMT) is an M-Estimate in which the function PSI is:

$$\text{PSI}(x) = \begin{cases} \text{Sin}(x/2.1) & |x| < 2.1\pi \\ 0 & \text{otherwise} \end{cases}$$

The fourth scheme then is to use the AMT estimator on the positive and negative pseudovalues separately, obtaining two estimates of ability. These two estimates are then weighted by the number of observations that went into them and summed. The resulting value is then divided by the total number of items and the result is the AMTed Jackknifed estimate.

We expect that when the test response pattern is reasonable (i.e. no responses are obtained which are unlikely based upon the Rasch Model), the AMT-Jackknife will look like the normal Jackknife. But when there are some odd

responses, they will not be counted as heavily, and so produce an estimate which is less affected by guessing and sleeping, while retaining the Jackknife's narrow sampling distribution.

Scheme 5 - WIM

Wright and Mead (Reference Note 4) developed a method for estimating ability in the Rasch model based upon an analysis of the residuals. Their method obtains an initial estimate of ability from raw score and it's associated standard error. Then it calculates the residual of each item's response for that person by subtracting from the response the probability of it being correct. These residuals are standardized and a t-statistic calculated for the fit of this person's response pattern. If this t is greater than some chosen value (say t=2), then all items more than two logits above the person's initial ability estimate are omitted from that person's test and a new ability estimate is obtained based upon the shortened test. This process is repeated until an acceptable t is achieved or until the test gets too short to work with.

This estimation scheme (WIM) was also included in our tests. The subroutine which does WIM estimation was written by Ronald Mead. Our results with this method reflect only on the method as we received it. We did not try to tune it by varying the critical t-value. It could be that it's performance would improve with fine tuning.

IV. The Guessing Model

How one characterizes individual responses in a simulation is critically important to its outcome. Certainly if one built an estimator that matched the response generator that estimator should win hands down in any competition. The validity of such investigations depends upon how the response model matches reality. We decided that a reasonable model for responding has the following characteristics:

1) Need - A person guesses if he/she has a need to guess. This is a function of the extent to which the item is more difficult than the person is able. If someone thinks they know answer they will not guess, if they don't they might.

2) Invitation - this is a function of the item unrelated to its difficulty (usually a function of the distractors). Some items invite one to guess -- others discourage it.

3) Inclination - A function of people unrelated to ability. Some people like to guess (risk takers?) and others do not (risk avoiders?).

4) Glitch - This represents something unexpected that may be an item-person interaction unrelated to ability, difficulty, inclination or invitation, a way for the best laid plans to go wrong.

The guessing model is:

$$\pi_{ij}=P_{ij}+(1-P_{ij})(V_j+C_i-V_jC_i)/u_j$$

where,

π_{ij} is the probability of person i getting item j correct

P_{ij} is the probability of person i getting item j correct based upon the Rasch model which is given earlier. The need to guess arises when P_{ij} is small because d_j is larger than a_i.

V_j is the invitation to guess associated with item j ($0 \leq V_j \leq 1$)

C_i is the inclination to guess associated with person i ($0 \leq C_i \leq 1$)

u_j is the number of alternatives for item j.

The actual response that was generated by this model was determined and it was allowed to remain with probability 1-G (where G is the Glitch factor) and was changed with probability G, the generating parameter included to stir up trouble and add noise.

V. The Simulation

There are a large number of things to vary in a simulation to get a full picture of what is happening. This simulation had eight factors which were systematically varied and on which all five estimation schemes were tried out. These were:

1) Difficulty distribution (3 levels) - There were three distributions of difficulties that were used; uniform, Gaussian and bimodal. The bimodal distribution was generated by constructing a uniform distribution and leaving out the middle half.

2) Test length (3 levels) - We simulated tests of three lengths, short (10 items), medium (20 items) and longish (40 items). Longer tests were not used because the generalizability of results would increase only slightly but computer costs would multiply.

3) Test width (2 levels) - Two test widths were simulated, narrow (2 logits) and medium (4 logits). Wider tests are in use, but that aspect must be left for another day.

4) Number of alternatives (two levels) - Tests with five choices were simulated since that reflects a common test format, as were tests with two alternatives (true-false format) which represents an extreme case.

5) Ability (4 levels) - Four levels of ability

were used; Extra Low, Low, Medium, and High. Typically we chose as Extra Low an ability that was the same as the easiest item on the test. Medium was typically chosen as zero, with Low halfway between them. High was usually symmetric with Low. Therefore with the difficulties shown previously the four abilities chosen would be -2, -1, 0, +1. There was some variation in this choice which will be explained later.

6) Invitation to guess (3 levels) - this ranged from low (0.0) to medium (0.5) to high (0.9).

7) Inclination to guess (3 levels) - The same as invitation. As is evident from the response model these two parameters are symmetric in their effect and so only the six interesting combinations were used.

8) Glitch (3 levels) - Glitch is meant to convey rare, or at most, seldom trouble. Thus we used three levels of glitch, none (0.0), a little (0.1), and a lot (0.4). Note that a glitch of 0.5 is maximum, in that it will make the expected score for any response pattern the same (L/2).

The Dependent Variables of the Simulation

Two aspects of estimator performance are of interest. The first is accuracy -- how different is the estimate of ability obtained from each estimator from the ability parameter which generated the response vector. We have summarized this by the mean difference between estimated ability for each estimator and the generating parameter. In the course of the simulation this was sometimes violated because as a response vector was generated it was checked to see if it was estimable. In particular if a response vector had a raw score of 1 or lower or L-1 or higher it was not used and another was

generated. This resulted in a truncation of the ability distribution. This truncation caused the low ability groups to have a somewhat higher ability than the generating parameter would indicate, and the High ability group to have a slightly lower ability than the generating parameter. To correct for this we estimated the Rasch ability without any noise for a particular simulation situation (a specific length, width, distribution and glitch) and used the pure Rasch ability estimates as the basis of comparison for that simulation. Hence when there is no noise the Rasch estimates have zero bias by construction.

The second aspect of estimator performance that interests us is the variance of the sampling distribution of that estimator around it's own mean. Of course the smaller this is the better the estimator.

We combined these two measures of estimator performance into a total variance figure by adding together the weighted squared bias (analogous to the between sum of squares) to the sampling variance (the within sum of squares) using the usual synthesis of variance weightings. This represents the overall efficiency of each estimator. We then found that estimator which had the smallest efficiency for that sample and divided each estimator's efficiency into it to obtain relative efficiency. It is this figure that we shall report.

VI. Results and Discussion

Obviously, with a design consisting of almost 4,000 cells and five estimators per cell it would be impractical

to attempt to present all the results. Instead we shall present selected findings representative of the main effects, and comment on some important interactions and trends. The principle effect is that there was a real winner -- the AMT-Jackknife. The AMT-Jackknife won not because it was the most bias free, although it did reasonably well in that regard, but rather because of its extremely small sampling variance.

No Noise

Before going on to the noisey simulations let us consider the uncontaminated situation. It would seem that any estimation scheme proposed must do reasonably well in this situation before it can be considered a viable alternative to ordinary methods.

Table 2 shows the relative efficiencies of the five estimators for three test lengths, two different widths and four abilities. These are rounded to one decimal place for usefulness.

Insert Table 2 About Here

The results for a uniform distribution of difficulties are striking for two reasons. First, the superiority of the AMT-Jackknife (followed closely by the standard Jackknife) is evident. This assures us that the Jackknife is a viable scheme. The second observation leads us to check the FORTRAN code. The Rasch maximum likelihood estimator is not

the most efficient. This counters expectation since maximum likelihood is supposed to yield estimates with minimum variance. Why does that fail to happen in this case? The answer is that the superlative properties of maximum likelihood are asymptotic. As test length increases the relative efficiency of the Rasch estimator goes up from 70% to 90%. The WIM estimator behaves in the same way. It would seem that 40 items is not enough for asymptotic properties to triumph over Jackknife properties. This finding leads us to reconsider using maximum likelihood with short tests without further thought. Replacing maximum likelihood with AMT-Jackknife may benefit short test applications. We are not the first to observe that maximum likelihood does not accomplish everything one would desire from efficient estimation. Lewis (1970) in studying methods for the estimation of thresholds of sensitivity curves (a problem similar to the one we are examining) found that maximum likelihood was unsatisfactory and used instead a scheme based on order statistics (The "Countback Method").

Insert Tables 3 and 4 About Here

Continuing to explore the efficiency of these five estimators in the errorless situation we see (Table 3) that the same structure observed for a uniform distribution holds for a Gaussian distribution. Once again the AMT-Jackknife is the winner, followed closely by the standard Jackknife

and then Rasch and WIM. In all situations the Traditional guessing correction is an abject failure. This is not unanticipated since it is making corrections for a disturbance that is totally absent. As we will see later, its performance improves when guessing does occur (not surprisingly). Incidently, WIM, which is the most computationally expensive procedure, is especially expensive for Gaussian and Bimodal distributions of difficulty. More iterations are required for convergence in these situations than when the difficulties are uniform.

Table 4 shows the efficiencies for a bimodal distribution with essentially the same structure evident that appeared with the other two distributions. WIM estimates were not obtained for a 40 item test (width 2) when the procedure had not converged after 100 seconds (on an Amdahl /V6). It was felt that any information obtained from such a result would not be worth the cost/effort.

One conclusion is clear; when there is no guessing, we can improve on the maximum likelihood estimator of ability in the Rasch model for tests of modest length (less than 40 items or so). In this noiseless situation there is little to choose between the robust AMT-Jackknife and the standard Jackknife. The AMT is a bit better, but uses a bit more effort in its computation. We also found that the traditional correction for guessing if applied when guessing is absent can have disastrous effects upon the efficiency of estimation. WIM, works as well as straight Rasch estimation

when there is no guessing, although it does lead to a bit of shrinkage due to the shortening of tests when unusual residuals occur by chance.

Some Guessing

The next step in the exploration of estimators of ability is to study their behavior with a little bit of noise. Tables 5, 6 and 7 show the relative efficiencies for the three distributions with guessing invitations and guessing inclinations set at 0.5. Even a cursory examination shows that the structure observed in the no noise situation still obtains. The AMT-Jackknife and the standard Jackknife still lead, but WIM and Traditional corrections are gaining. The bimodal distribution seems to trouble the Jackknife more than its robustified version, but both seem to do tolerably well. As one would suspect, at lower ability levels schemes which are designed to deal with guessing (WIM and Traditional) work to their best advantage. At higher ability levels this is not the case. Jackknifing schemes do better on narrow tests than wide ones (this observations has been confirmed by examining their behavior on very wide tests of 6 to 8 logits and noting a deterioration of performance. This is especially marked on 8 logit wide tests for the AMT).

Insert Tables 5, 6, and 7 About Here

The conclusions reached for noiseless data still hold,

but less strongly. The two Jackknife methods remain the schemes of choice, especially for ability individuals above the mean. But as the data get increasingly noisey each estimator reacts in its own way. The Rasch estimator yields the same score for all raw scores of the same value, regardless of how that raw score was obtained, but indicates its displeasure by yielding a poor goodness-of-fit statistic for misfitting persons. WIM reacts by shortening the test, telling us in essence that only a small portion of the test response vector obeys the Rasch model. The Jackknife methods react by regressing the scores toward zero (increasing bias but reducing variance of sampling distribution) but increasing the standard error. Thus saying that the information on this individual is small.

More Guessing

Let us continue to follow the pattern by considering the same three distributions of item difficulty, but this time with a great deal of guessing. Tables 8, 9, and 10 show the results when guessing invitation and inclination are both set to 0.9. This yields a situation in which a person guesses whenever he doesn't know the answer and is identical to the situation posited in the derivation of the Traditional guessing correction. In this situation we would expect the Traditional method to shine and it does do well, but only when the test length is great enough to overcome its small sample inefficiency.

Insert Tables 8, 9 and 10 About Here

Once again the same pattern of results emerges. For short tests the Jackknifing schemes work best, with the edge always in the direction of the AMT. As tests get longer (40 items) the Traditional guessing correction starts to work quite well. WIM on the other hand is disappointing, doing scarcely better than just a straight Rasch estimate. This must be interpreted, however. WIM reduces measurement bias quite well. But in doing so it also decreases test length substantially, one could argue that the length of test evaluated by WIM, after eliminating items with large residuals, corresponds to the test that the testee actually took. However, the reduced test length has the concomitant effect of increasing the standard error of measurement, and this causes its disappointing showing in the efficiency statistic.

Guessing + Glitching

Since the distribution of difficulties does not appear to have much effect on the behavior of the various estimators, we shall confine the remainder of the results we report to one or the other of the distributions, with only side comments if the results differ substantially when another distribution is used. (Incidentially for an extremely bimodal distribution in which all items are piled up at the extremes, the AMT will not work at all).

Insert Table 11 About Here

Table 11 shows the reaction of the various estimators to Glitch of 0.1 over several test widths and for different amounts of guessing. There are no surprises. The deterioration of performance of the Jackknifing estimators with increased test width is visible but not too bad. The AMT is always superior to the standard Jackknife. Under all conditions Jackknifing seems to be the best choice for higher ability individuals. Jackknifing also works rather well for correcting guessers, but there the other methods may be better. We have only reported results for tests of length 20 in this Table, but this is representative of the general findings. The Jackknifing methods do relatively less well with a test of length 40, and do relatively better with a test of length 10.

True/False Tests

If the number of alternatives is shrunk from five to two we find much the same results. With no guessing the Jackknifing methods do best with an edge to the AMT. As guessing gets increasingly prevalent the Traditional correction scheme works better. But we still find that for high abilities the AMT method is superior in efficiency to all others.

VII. Standard Errors

The Rasch standard error is,

$$\text{Rasch(s.e.)} = 1/\text{SQRT}\{\text{SUM}_j[P_{ij}(1-P_{ij})]\}$$

for each ability level i. This accurately reflects what was observed empirically for the Rasch ability estimates in our simulations. The standard deviations of the sampling distributions, when there was no guessing, was about what this equation would predict. It underpredicted the variability observed when there was noise. The WIM standard error is calculated in the same way as the Rasch, except for a test of reduced length. This seems to accurately reflect reality for the situations we tested.

The Jackknife standard error is calculated directly from the pseudovalues by:

$$\text{Jackknife(s.e.)} = \text{SQRT}[\text{SUM}_j\langle a_j^* - a^*\rangle^2/\langle (L-1)L\rangle]$$

and is known to be a conservative estimator. This is certainly true in this case. It tends to overestimate the actual s.e. by about 50% for test lengths of 10, by 25% for test lengths of 20. But it is just about right for test lengths of 40.

There are several candidates for estimating the standard error of the AMT, but our investigations are insufficient to be able to recommend one at this time. It seems reasonable to use the corrected Jackknife standard error until a better choice is found. The Jackknife s.e. will almost certainly be conservatively large.

VIII. Conclusions

This investigation sought to find and test alternative methods for estimating ability under the Rasch model in the

face of plausible noise. We did this by using some recent developments in robust estimation, without adding parameters to the model and so retained the Rasch model's attractive attributes. In our investigation we found that gains in recovering abilities in the presence of guessing and untoward responses of other kinds can be obtained through the use of a robustified Jackknife. But we also found that specially developed models aimed at the lower end of the ability continuum may be able to accomplish this better than these general tools. WIM worked when there was guessing, and aided in increasing the accuracy of estimation for low ability testees. The Traditional method worked when there was a lot of guessing, the test was long, and the ability of the testees was low.

A surprising finding was that for short tests of 10 or 20 items the Jackknife estimators, with a significant edge to the AMT version, yielded better estimates of ability than the maximum likelihood estimator even when pre-conditions for the Rasch model held. This increase in efficiency of estimation is especially important for those applications of latent trait models which use a limited number of measures obtained about a person as a de facto "test" (see for example the analysis of parole data in Perline, Wright and Wainer, 1979). In these circumstances the number of items cannot be increased sensibly and the only alternative is to improve the estimate of ability through other means. Thissen (1976) attempted to do this by using a method Bock (1972)

developed on the wrong answers, but this is very expensive computationally and only applicable to multiple choice items. Super efficient estimators may also be useful in such applications as adaptive testing.

The simulations we did were very extensive, nevertheless they barely made a dent in what needs to be done. A careful study of estimators of standard error is critical, as are the distributional properties of the Jackknifed estimators. To our knowledge, no one has used robust estimators in conjunction with the Jackknife before and so nothing is known about that distribution. We believe Jackknife estimates are t-distributed (although there is difficulty in determining the effective degrees of freedom). It seems reasonable therefore to suppose that the robust Jackknife will have a similar symmetric (albeit tighter) distribution. This suggests that the Jackknife estimates of standard error for the AMT estimator are conservative. Just how conservative these actually are however awaits further investigation.

A second area of investigation that is still incomplete are goodness-of-fit tests. Substituting robust estimates of ability into the usual goodness-of-fit equations should yield a conservative estimate more realistic than those usually obtained (which benefit from capitalization on chance). But we do not know to what extent the asymptotic properties of such fit statistics derived and/or described by Andersen (1973), Fischer (1974), Martin-Lof (1974), and

Wright and Stone (1979) apply.

The finding of improved estimation efficiency is an intriguing one. Lewis(1970) pointed out that although maximum likelihood estimates of location parameters of ogive functions are asymptotically identical to minimum chi-square estimates, they can be quite different for small samples. Neither makes any claims for small sample efficacy, but what is surprising is how large "small" can be, and how much of an improvement can be made using an alternative procedure. Lewis found that asymptotically optimal procedures did especially badly in estimating accurate confidence intervals around the location parameter. Perhaps this too is an area in which the AMT-Jackknife will prove useful. The questions are clear and important, and the methodology for answering them is straightforward.

There are a number of other estimators which may improve performance still more. For example, Ramsay (1977) found that the E_a estimator has some advantages over the AMT. Novick (Reference Note 2) has suggested several Bayesian estimators that may have promise.

The key point of this paper is that for short tests the asymptotic properties of maximum likelihood estimators are not fully realized. Other methods increase efficiency. In addition, these other estimators can correct for noise in the data like guessing and so can increase validity. The AMT-Jackknife may not be the best estimator of its type that can be derived. Perhaps other variations on this theme can

go even further in the direction of super-efficiency. Nevertheless the AMT-Jackknife does seem to deal well with the problem of guessing that is so poorly handled by trying to estimate a lower asymptote of the item characteristic curve.

Reference Notes

(1) Lord, F. (1979) "Small N Justifies Rasch Methods". Talk given at the 3rd Conference on Computerized Adaptive Testing. Minneapolis, Minnesota.

(2) Novick, M. R. (1979) Informal discussion after talk on robust estimation. The Third Conference on Computerized Adaptive Testing. Minneapolis, Minnesota.

(3) Ree, M. J. and Jensen, H. E. (1979) "Effects of Sample Size in the Estimation of Item Parameters" . Talk given at the 3rd Conference on Computerized Adaptive Testing. Minneapolis, Minnesota.

(4) Wright, B. D. and R. J. Mead (1976) "Analysis of residuals", unpublished paper.

References

Andersen, E. B. (1973) A goodness of fit test for the Rasch model. Psychometrika , 38, 123-140.

Andrews, D.F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972) Robust Estimates of Location, Princeton, N.J.:Princeton University Press.

Bock, R. D. (1972) Estimating item parameters and latent trait ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock, R. D. and Wood, R. (1971) Test Theory, Annual Review of Psychology, 22, 193-224.

Fischer, G. H. (1974) Einfurung in die Theorie psychologischer Tests. Grundlagen und Anwendungen. Bern: Huber.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., and Gifford, J. A. (1978) Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48, 467-510.

Lewis, C. (1970) The countback method. Research Bulletin 70-30. Princeton, N.J.: Educational Testing Service.

Lord, F. M. and Novick, M. R. (1968) Statistical Theories of Mental Test Scores. Addison-Wesley: Reading, Mass.

Martin-Lof, P. (1974) Exact tests, confidence regions and estimates. Proceedings of conference on fundamental questions in statistical inference. Memoirs, No. 1, Department of Theoretical Statistics, Institute of Mathematics, University of Aaarhus.

Perline, R., Wright, B. D., and Wainer, H. (1979) The Rasch model as additive conjoint measurement, Applied Psychological Measurement, 3, xx-xx.

Quenouille, M. (1956) Notes on bias in estimation, Biometrika, 43, 353-360.

Rasch, G. (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.

Ramsay, J. O. (1977) A comparative study of several robust estimates of slope, intercept, and scale in linear regression. Journal of the American Statistical Association, 72, 608-615.

Thissen, D. M. (1976) Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 13, 201-214.

Tukey, J. W. (1958) Bias and confidence in not quite large samples, (abstract), Annals of Mathematical Statistics, 29, 614.

Tukey, J. W. (1977) Exploratory Data Analysis, Addison-Wesley: Reading, Mass.

Wainer, H., Morgan, A, and Gustafsson, J-E (1979) A review of estimations procedures for the Rasch model with an eye toward longish tests. Under review.

Wright, B. D. and Panchapakesan, N. (1969) A procedure for sample free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. (1968) Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 85-101.

Wright, B. D. (1977) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D. and Mead, R. J. (1977) The Use of Measurement Models in the Definition and Application of Social Science Variables. Arlington, VA: U.S. Army Research Institute Technical Report DAHC19-76-G-0011.

Wright, B. D. and Stone, M. H. (1979) Best Test Design, MESA Press: Chicago.

TABLE I

THE ABILITY-RAW SCORE CONVERSION MATRIX

Raw Score	Ability All Items	Ability Omitting Item (I)									
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	-2.78	-2.32	-2.45	-2.56	-2.63	-2.68	-2.72	-2.74	-2.75	-2.76	-2.77
2	-1.83	-1.37	-1.45	-1.54	-1.63	-1.69	-1.74	-1.77	-1.79	-1.80	-1.81
3	-1.15	-0.68	-0.73	-0.80	-0.88	-0.95	-1.01	-1.05	-1.09	-1.11	-1.12
4	-0.56	-0.07	-0.10	-0.15	-0.21	-0.29	-0.35	-0.41	-0.46	-0.49	-0.51
5	0.00	0.51	0.49	0.46	0.41	0.35	0.29	0.21	0.15	0.10	0.07
6	0.56	1.12	1.11	1.09	1.05	1.01	0.95	0.88	0.80	0.73	0.68
7	1.15	1.81	1.80	1.79	1.77	1.74	1.69	1.63	1.54	1.45	1.37
8	1.83	2.77	2.76	2.75	2.74	2.72	2.68	2.63	2.56	2.45	2.32
9	2.78										

TABLE 2

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = 0
GUESSING INCLINATION = 0
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A UNIFORM DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.7	.7	.7	.7	.8	.8	.9	.8	.9	.9	.9	.9
	Traditional	.2	.1	.2	.2	.1	.1	.2	.4	.0	.1	.2	.3
	Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	WIM	.7	.7	.7	.7	.8	.9	.9	.8	.9	.9	.9	.9
4	Rasch	.8	.7	.7	.8	.8	.8	.8	.8	.9	.9	.9	.9
	Traditional	.2	.2	.2	.3	.1	.1	.2	.4	.0	.0	.2	.3
	Jackknife	1.0	.9	.9	1.0	1.0	.9	.9	.9	1.0	1.0	.9	.9
	AMT	.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	WIM	.7	.7	.6	.7	.7	.7	.8	.7	.7	.9	.9	.9

TABLE 3

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = 0
GUESSING INCLINATION = 0
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A GAUSSIAN DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.7	.7	.7	.7	.8	.9	.8	.8	.9	.9	.9	.9
	Traditional	.2	.1	.2	.2	.1	.1	.2	.3	.0	.1	.2	.3
	Jackknife	1.0	1.0	.9	.9	1.0	1.0	.9	1.0	1.0	1.0	1.0	1.0
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	WIM	.7	.7	.7	.7	.8	.8	.8	.8	.9	.9	.9	.9
4	Rasch	.8	.7	.7	.7	.8	.8	.8	.8	.8	.9	.8	.8
	Traditional	.1	.2	.2	.3	.1	.1	.2	.4	.0	.0	.2	.3
	Jackknife	1.0	1.0	.8	.9	1.0	.9	.9	.9	1.0	1.0	.9	.9
	AMT	.8	1.0	1.0	1.0	.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	WIM	.7	.7	.6	.6	.8	.6	.8	.7	.9	.8	.9	.8

TABLE 4

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = 0
GUESSING INCLINATION = 0
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A BIMODAL DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.6	.6	.5	.6	.8	.7	.6	.6	.9	.9	.6	.9
	Traditional	.1	.1	.1	.2	.1	.1	.1	.2	.0	.1	.1	.2
	Jackknife	.8	.7	.6	.8	1.0	.7	.6	.7	1.0	1.0	.6	1.0
	AMT	1.0	1.0	1.0	1.0	.8	1.0	1.0	1.0	.8	1.0	1.0	1.0
	WIM	.6	.6	.5	.6	.8	.6	.6	.6	.9	.9	.6	.9
4	Rasch	.8	.6	.2	.8	.8	.9	.2	.9	.9	1.0	.2	.9
	Traditional	.2	.1	.0	.2	.1	.1	.0	.3	.0	.1	.0	.2
	Jackknife	1.0	.7	.2	.9	1.0	1.0	.2	1.0	1.0	1.0	.2	1.0
	AMT	.6	1.0	1.0	1.0	.3	.7	1.0	.9	.2	.4	1.0	.5
	WIM	.7	.6	.2	.7	.8	.8	.2	.8	*	*	*	*

TABLE 5

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = .5
GUESSING INCLINATION = .5
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEMS DIFFICULTIES HAVE A UNIFORM DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.8	.8	.7	.6	1.0	.9	.9	.8	1.0	1.0	.9	.8
	Traditional	.2	.3	.3	.4	.4	.5	.5	.5	.5	1.0	.7	1.0
	Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.9
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	WIM	.8	.8	.7	.6	1.0	.9	.9	.8	1.0	1.0	.9	.8
4	Rasch	.9	.8	.7	.6	.9	1.0	.8	.8	.8	1.0	.9	.7
	Traditional	.4	.5	.3	.4	.6	.6	.7	.5	.4	1.0	.7	.8
	Jackknife	1.0	.9	.8	.9	.9	1.0	.9	.8	.8	1.0	.9	.8
	AMT	1.0	1.0	1.0	1.0	.9	1.0	1.0	1.0	.8	.9	1.0	1.0
	WIM	.8	.8	.6	.5	1.0	1.0	.8	.6	1.0	1.0	.9	.7

TABLE 6

RELATIVE EFFICIENCES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = .5
GUESSING INCLINATION = .5
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A GAUSSIAN DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.8	.8	.7	.6	1.0	.9	.8	.7	1.0	.9	.8	.8
	Traditional	.3	.3	.4	.4	.4	.4	.6	.5	.5	1.0	.7	1.0
	Jackknife	1.0	1.0	1.0	.9	1.0	1.0	.9	.9	1.0	.9	.9	.8
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	WIM	.8	.8	.7	.6	1.0	.9	.8	.7	1.0	.9	.9	.8
4	Rasch	1.0	.9	.7	.5	1.0	1.0	.7	.6	1.0	.8	.8	.8
	Traditional	.4	.5	.4	.3	.6	.7	.6	.4	.5	1.0	.7	.8
	Jackknife	1.0	1.0	.9	.8	.9	1.0	.8	.8	1.0	.8	.8	.8
	AMT	.9	1.0	1.0	1.0	.9	1.0	1.0	1.0	.9	.8	1.0	1.0
	WIM	.8	.8	.7	.5	1.0	1.0	.7	.6	1.0	.9	.9	.8

TABLE 7

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = .5
GUESSING INCLINATION = .5
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A BIMODAL DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.8	.6	.5	.5	1.0	.8	.6	.5	1.0	.8	.6	.4
	Traditional	.3	.3	.3	.2	.6	.6	.4	.3	.4	1.0	.4	.6
	Jackknife	1.0	.7	.7	.7	1.0	.8	.6	.5	1.0	.8	.6	.5
	AMT	1.0	1.0	1.0	1.0	.9	1.0	1.0	1.0	.8	.8	1.0	1.0
	WIM	.8	.6	.5	.5	.9	.8	.6	.5	1.0	.8	.6	.4
4	Rasch	.9	.6	.2	.7	.7	1.0	.2	.8	1.0	.8	.2	.5
	Traditional	.4	.4	.1	.4	.4	.8	.2	.5	.6	1.0	.1	.6
	Jackknife	1.0	.6	.2	.9	.7	1.0	.2	1.0	1.0	.8	.2	.5
	AMT	.6	1.0	1.0	1.0	.4	.9	1.0	1.0	.4	.6	1.0	1.0
	WIM	.8	.4	.2	.5	1.0	1.0	.2	.7	*	*	*	*

TABLE 8

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = .9
GUESSING INCLINATION = .9
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEMS DIFFICULTIES HAVE A UNIFORM DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.8	.8	.7	.6	1.0	.9	.8	.8	.7	.5	.8	.6
	Traditional	.4	.4	.4	.5	.7	.9	1.0	.8	1.0	1.0	1.0	1.0
	Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	.9	1.0	.7	.5	.8	.7
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	.9	1.0	.7	.5	.8	.7
	WIM	.8	.8	.7	.6	1.0	.9	.8	.8	.7	.5	.8	.6
4	Rasch	1.0	.9	.7	.6	.9	1.0	.6	.6	.6	.5	.7	.7
	Traditional	.6	.8	.5	.4	.9	.9	1.0	.6	1.0	1.0	1.0	1.0
	Jackknife	1.0	1.0	.8	.9	.8	1.0	.7	.7	.6	.5	.7	.8
	AMT	1.0	1.0	1.0	1.0	.8	1.0	.8	1.0	.5	.5	.8	1.0
	WIM	.8	.8	.6	.5	1.0	.9	.6	.5	.7	.6	.7	.6

TABLE 9

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = .9
GUESSING INCLINATION = .9
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A GAUSSIAN DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.9	.8	.7	.5	1.0	.9	.8	.7	.7	.5	.7	.6
	Traditional	.5	.4	.4	.4	.6	.8	.9	.7	1.0	1.0	1.0	1.0
	Jackknife	1.0	1.0	.9	.9	1.0	1.0	.9	.9	.7	.5	.8	.7
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.7	.6	.8	.8
	WIM	.9	.8	.7	.5	1.0	.9	.8	.7	.7	.5	.7	.6
4	Rasch	1.0	.9	.6	.4	1.0	1.0	.6	.5	.6	.5	.6	.6
	Traditional	.6	.9	.5	.4	1.0	.8	1.0	.5	1.0	1.0	1.0	.8
	Jackknife	1.0	1.0	.8	.8	.9	1.0	.7	.7	.6	.5	.7	.6
	AMT	1.0	1.0	1.0	1.0	.9	1.0	.8	1.0	.6	.5	.8	1.0
	WIM	1.0	.8	.5	.4	1.0	.9	.6	.4	1.0	.9	.7	.5

TABLE 10

RELATIVE EFFICIENCIES OF 5 ESTIMATORS ON TESTS OF VARIOUS LENGTHS AND WIDTHS

GUESSING INVITATION = .9
GUESSING INCLINATION = .9
GLITCH = 0

ITEMS HAVE 5 CHOICES
ITEM DIFFICULTIES HAVE A BIMODAL DISTRIBUTION

WIDTH (Logits)		LENGTH (Number of Items)											
		10				20				40			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
2	Rasch	.7	.6	.5	.4	1.0	.9	.6	.5	.6	.5	.6	.5
	Traditional	.4	.4	.2	.4	.8	.8	.8	.4	1.0	1.0	1.0	.9
	Jackknife	.9	.7	.6	.6	1.0	.9	.7	.6	.6	.5	.7	.6
	AMT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.6	.6	1.0	1.0
	WIM	.7	.6	.5	.4	.9	.9	.6	.5	.6	.5	.6	.5
.4	Rasch	1.0	.6	.2	.4	.6	.9	.2	.5	.5	.4	.2	.3
	Traditional	.9	.5	.2	.4	.7	1.0	.4	.6	1.0	1.0	.4	.5
	Jackknife	1.0	.6	.3	.6	.6	.9	.2	.6	.5	.4	.2	.3
	AMT	.9	1.0	1.0	1.0	.4	1.0	1.0	1.0	.2	.4	1.0	1.0
	WIM	1.0	.5	.2	.4	1.0	.9	.2	.5	*	*	*	*

TABLE 11

RELATIVE EFFICIENCIES OF VARIOUS ESTIMATORS OF ABILITY FOR A TEST WITH 20 ITEMS
WHOSE DIFFICULTIES ARE UNIFORMLY DISTRIBUTED
THERE IS A RANDOM NOISE COMPONENT OF 10% (GLITCH = .1)
100 ENTRIES SAMPLED PER CELL IN DESIGN

Amount of Guessing (V, C)	Estimator	WIDTH (Logits)											
		2				4				6			
		Ability				Ability				Ability			
		X. Low	Low	Med.	High	X. Low	Low	Med.	High	X. Low	Low	Med.	High
(0,0)	Rasch	1.0	.9	.9	.8	1.0	.8	.8	.9	.7	.9	.5	.9
	Traditional	.2	.1	.2	.3	.4	.1	.2	.3	.8	.2	.1	.3
	Jackknife	1.0	1.0	1.0	1.0	.9	.9	.9	1.0	.6	1.0	.6	1.0
	AMT	1.0	1.0	1.0	1.0	.9	1.0	1.0	1.0	.4	.9	1.0	.9
	WIM	1.0	.9	.9	.8	1.0	.6	.8	.8	1.0	.7	.3	.7
(.5,.5)	Rasch	1.0	.9	.9	.8	.6	1.0	.8	.8	.4	.9	.6	.8
	Traditional	.9	.5	.6	.4	1.0	.6	.6	.3	1.0	.7	.4	.4
	Jackknife	1.0	1.0	1.0	1.0	.6	1.0	.9	.9	.4	.9	.6	1.0
	AMT	1.0	1.0	1.0	1.0	.6	1.0	1.0	1.0	.3	.8	1.0	1.0
	WIM	1.0	.9	.9	.8	.7	1.0	.8	.6	.8	1.0	.5	.6
(.9,.9)	Rasch	1.0	.9	.9	.8	.7	.9	.8	.7	.3	.7	.6	.9
	Traditional	.8	1.0	.7	.5	1.0	1.0	.7	.4	1.0	1.0	.5	.3
	Jackknife	1.0	1.0	1.0	1.0	.7	.9	.9	.9	.3	.7	.6	1.0
	AMT	1.0	1.0	1.0	1.0	.6	1.0	1.0	1.0	.3	.7	1.0	.8
	WIM	1.0	.9	.9	.7	.8	1.0	.8	.8	.7	.9	.4	.6